December 13, 2023

E-Filed

The Honorable Donna M. Ryu
United States District Court for the Northern District of California
Oakland Courthouse, Courtroom 4 – 3rd Floor
1301 Clay Street
Oakland, California 94612

Re: *Doe 1, et al. v. GitHub, Inc., et al.;* Lead Case No. 4:22-cv-06823-JST

Dear Judge Ryu:

Pursuant to the Court's Order Referring All Discovery Disputes to Magistrate Judge (ECF No. 175) and Paragraph 14 of this Court's Standing Order, Plaintiffs Does 1–5 ("Plaintiffs") submit this letter brief regarding Microsoft Corporation's ("Microsoft") responses to Plaintiffs' First Set of Interrogatories propounded on July 19, 2023 ("Interrogatories"). Plaintiffs seek an order compelling Microsoft to provide responses to Interrogatories 2, 6, 11, and 13.

Consistent with the Standing Order, Plaintiffs attempted to further meet and confer with Microsoft regarding its plainly inadequate discovery responses. Microsoft has refused to meet and confer in person. Plaintiffs have previously met and conferred by zoom on September 7, 2023. Plaintiffs subsequently confirmed the parties' positions in writing on September 19, 2023. Even after Your Honor's appointment, Microsoft has dragged its feet and avoided its discovery obligations. Plaintiffs have made efforts to submit a joint letter. Microsoft has refused. Plaintiffs provided a copy of this letter to Microsoft prior to filing. Microsoft's approach is inconsistent with Rules 26, 33, 37, as well as the local rules of the Court and the Standing Order.

Current Case Deadlines: (1) Fact Discovery Cut-Off: September 27, 2024; (2) Expert Discovery Cut-Off: February 21, 2025; (3) Dispositive Motion Hearing: not scheduled; (4) Class Certification Motion: March 27, 2025; (4) Pretrial Conference and Trial Dates: not scheduled.

Interrogatory 2 calls for the identities of persons who have managed and directed Microsoft during the relevant period. This is basic information, routinely provided in civil litigation. *See, e.g.*, *In re TFT-LCD (Flat Panel) Antitrust Litig.*, No. M 07-1827 SI, 2007 WL 2782951, at *2 (N.D. Cal. Sept. 25, 2007) (permitting interrogatories seeking the "names, positions, dates of employment/tenure, and addresses" of *inter alia*, directors and officers as an exception to a general discovery stay); *In re Folding Carton Antitrust Litig.*, 76 F.R.D. 417, 419 (N.D. Ill. 1977) (describing this information as "classic first-wave discovery"). Microsoft states it "will not undertake the burdensome investigation necessary to respond to this interrogatory" and has never offered any substantive response, Plaintiffs agree to narrow Interrogatory 2 to request names and dates in such positions of past and present officers and directors in one of those positions from 2017 on, including the period Microsoft negotiated and acquired GitHub.

Interrogatory 6: This interrogatory seeks names of individuals responsible for negotiating Microsoft's 2018 acquisition of GitHub for $7.5 billion. Microsoft and GitHub had close connections pre-acquisition.[1] GitHub's library of open-source code was used to train Copilot.

[1] *See* https://www.theverge.com/2018/6/4/17422788/microsoft-github-acquisition-official-deal/.

Microsoft purchased GitHub's library, including the repositories containing the code subject to the licenses at issue in the case, related products, and product conceptualizations. Responsive individuals—including GitHub's agents—will have relevant knowledge, including plans for products that became Copilot. They will know the basis for the $7.5 billion purchase price and can also reasonably be expected to know what, if any, plans or actions were taken regarding the open-source licenses, including related discussions of whether and how to ignore or violate them. This interrogatory merely seeks witness identities. Rule 33 requires the responding party to provide information using any information readily available to the responding party, including information known by employees, or any information contained in files maintained, or otherwise available. Again, Microsoft has not offered any response.

Interrogatory 11: This interrogatory seeks identification of software, databases, or services used to maintain, supervise, manage, analyze, program, update, troubleshoot, diagnose, test, or modify Copilot—basic information necessary to investigate and prove how Copilot works. Courts have ordered defendants to produce information that explains how a software product works. *See Facedouble, Inc. v. Face.com, Inc.*, No. 12CV1584-DMS MDD, 2014 WL 585868, at *2 (S.D. Cal. Feb. 13, 2014) (ordering production of "a guide or road map to [defendant's] source code," and recognizing difficulty in relying on examination of source code alone to determine how defendant's software works). This information will streamline future discovery by identifying potential sources of discoverable ESI and will inform decisions regarding expert hiring and use. Microsoft's response that it "provides hosting and technical support services to GitHub, which hosts Copilot on Azure servers. Microsoft's investigation is ongoing" is general, obfuscatory, and ducks the question. Microsoft should answer the question immediately. As a proposed compromise, Plaintiffs agree to narrow Interrogatory 11 to a request for a road map explaining how Copilot works with references to the names of software, databases, or services ***currently used*** to maintain, supervise, manage, analyze, program, update, troubleshoot, diagnose, test, or modify Copilot. *See Facedouble, Inc. v. Face.com, Inc.*, No. 12CV1584-DMS (MDD), 2014 WL 585868, at *2 (S.D. Cal. Feb. 13, 2014) (ordering production of "a guide or road map to [defendant's] source code. . . .").

Interrogatory No. 13: This interrogatory seeks information regarding the composition of the training data at issue in this case, how it was acquired, and relevant corporate policies—data used for a core issue in this case: training Copilot. *See, e.g.*, FAC ¶¶ 2, 10, 84–95, 128. Defendants have kept secret the full composition of training datasets for GPT-3, GPT-4, Codex, and Copilot. The most complete description of any of the training datasets is the general statement that Codex comes from GitHub, that it was trained on "billions of lines of source code from publicly available sources, including code in public GitHub repositories." FAC ¶ 87 (quoting https://github.blog/2021-06-30-github-copilot-research-recitation/). Courts routinely permit discovery into matters that allow Plaintiffs to ascertain the extent of copyright-related violations. *See, e.g.*, *Transglobal Commc'ns Grp., USA, Inc. v. Stone Sapphire, Ltd.*, No. CV 07-0500-GW(RCX), 2008 WL 11339591, at *2 (C.D. Cal. Mar. 21, 2008) (ordering defendants to respond to interrogatory which asked for information about factories defendants have used or use that "relates to plaintiff's allegation that defendants are manufacturing or producing products containing plaintiff's copyrighted works"). Courts also regularly order that parties produce information that explains how a software product works. *See, e.g.*, *Facedouble, Inc.*, 2014 WL 585868, at *2. The most likely source of proof supporting Plaintiffs' allegations involving training data is a description of the data used to train the models. A full response will also be relevant to Defendants' scienter in connection with removal of CMI. Plaintiffs' claims will require a showing CMI was included in training data when it was acquired by Defendants. Plaintiffs agree to accept a list of all materials used to train GPT-3, GPT-4, Codex, and Copilot.

/s/ Joseph R. Saveri
Joseph R. Saveri
Joseph Saveri Law Firm, LLP
Attorneys for Plaintiffs